Adding population structure to models of language evolution by iterated learning

Adding population structure to models of language evolution by iterated learning

Andrew Whalen

University of Edinburgh

Thomas L. Griffiths

University of California, Berkeley

Author Note

Word count: 4000
Address for correspondence:
Andrew Whalen,
The Roslin Institute,
University of Edinburgh,
Easter Bush,
Midlothian, UK, EH25 9RG.
Email: awhalen@roslin.ed.ac.uk

Abstract

Previous work on iterated learning, a standard language learning paradigm where a sequence of learners learns a language from a previous learner, has found that if learners use a form of Bayesian inference, then the distribution of languages in a population will come to reflect the prior distribution assumed by the learners (Griffiths and Kalish 2007). We expand these results to allow for more complex population structures, and demonstrate that for learners on undirected graphs the distribution of languages will also reflect the prior distribution. We then use techniques borrowed from statistical physics to obtain deeper insight into language evolution, finding that although population structure will not influence the probability that an individual speaks a given language, it will influence how likely neighbors are to speak the same language. These analyses lift a restrictive assumption of iterated learning, and suggest that experimental and mathematical findings using iterated learning may apply to a wider range of settings.

Adding population structure to models of language evolution by iterated learning

Language changes; English today is slightly different from a hundred years ago, and radically different from a thousand years ago. An important cause of language change is the variation that occurs during the language learning process (see, e.g., DeGraff, 2001). One of the major tools that has been used to study the impact of language learning on the structure of languages is the iterated learning model (Kirby, 2001). In iterated learning, a set of simulated learners each learn language from the utterances of other learners and then produce utterances themselves that are provided to other learners. Repeating this process, the learners reshape the language. Simple learning algorithms can lead to significant changes, increasing the regularity of languages (Kirby, 2001; Smith, Kirby, & Brighton, 2003; Brighton, 2002) and expressing or even emphasizing the biases of learners (Griffiths & Kalish, 2007; Kirby, Dowman, & Griffiths, 2007).

The simplest iterated learning model – the case that submits most easily to mathematical analysis – is the transmission chain, in which each learner learns from the previous learner and generates utterances for the next. However, more complex models are possible. Exploring these models is important in two ways. First, it lets us establish the generality of results obtained for transmission chains, which represent the majority of previous analyses. Second, it allows us to explore phenomena that only emerge in more complex models. For example, speakers of the same language tend to cluster together spatially – something that is hard to explain using transmission chains.

In this paper, we explore how more complex population structures influence the outcome of iterated learning. We begin by introducing a formal framework for analyzing iterated learning in which learning is modeled as Bayesian inference. We then build on previous analyses of transmission chains by Griffiths and Kalish (2007), showing that similar analytic results can be obtained with populations where the relationships between learners can be expressed as a heterogeneous graph. We verify these results using simulations with two-dimensional lattices, small-world graphs (Watts & Strogatz, 1998) and scale-free graphs (Barabasi & Albert, 1999), population structures that mimic some of the properties of real populations. These simulations

show that neighbors in a graph are more likely to share the same language than is expected by chance. To quantify this effect we utilize techniques developed for voter models (Sood, Antal, & Redner, 2008; Castellano, 2012) and show that although the graphical structure of a population does not change how likely a individual learner speaks a certain language, it does impact how likely it is that neighbors will be able to communicate.

Iterated Bayesian learning

In the simplest iterated learning model, a population is assumed to be a series of parallel transmission chains. At each step in the chain, a learner learns a language from a single teacher and then transmits a language to a single student. The dynamics of this process depend on the learning algorithm that is used by the students.

One way to specify a learning algorithm is to assume that learners use a form of Bayesian inference (Griffiths & Kalish, 2007). Adopting a language then becomes a statistical inference task where the inductive biases of learners – those factors other than the data that lead them to favor one language over another – are expressed as a prior probability distribution over languages. Under this assumption, learners choose to speak a language, L, based on hearing linguistic data, D. We assume that the probability of speaking L is the same as the posterior probability of the language, calculated using Bayes' rule,

$$p(L|D) = \frac{p(D|L)p(L)}{p(D)}, \tag{1}$$

where $p(L)$ is the prior probability of the language, which may not be equal across languages.

Griffiths and Kalish (2007) showed that for transmission chains the probability that a learner speaks a language, L, after a large number of generations is the same as the prior probability of the language, $p(L)$. Formally, the stationary distribution of the resulting stochastic process is the prior distribution over languages. This result is interesting because it suggests that the variation observed in modern languages can be directly connected to the inductive biases of human language learners. Kirby et al. (2007) expanded on this result, showing that variations on

Bayesian learning in which learners are more likely to choose languages with higher posterior probabilities can exaggerate the impact of the prior on the stationary distribution, allowing weak inductive biases to have a strong effect on the structure of the languages produced by iterated learning.

However, this simplest iterated learning model may not accurately represent real populations. To explore the generality of these results, Smith (2009) relaxed the assumption of learning from a single teacher and examined populations of learners who learned a single language from multiple teachers. Using simulations, Smith showed that the language such learners acquire is highly dependent on the initial distribution of languages in a population, and more weakly influenced by prior probabilities. Burkett and Griffiths (2010) pursued these results further, and found that if learners could learn multiple languages from multiple teachers, the distribution of languages in the population over a number of generations will still mirror the prior probability of each language. Convergence to a stable equilibrium that is not the prior distribution can also occur if fitness is added into the model (Kalish, 2007).

In the remainder of the paper, we relax a different assumption and consider learners in a structured population who each learn from a single teacher. The goal of this model is to examine whether the structure of a population will affect the long-term distribution of languages in the population.

Introducing population structure

A natural way to capture population structure in cultural evolution is to analyze evolutionary dynamics on graphs, where each node is an agent and edges indicate connections between those agents (e.g., Nowak, 2006). In this section, we analyze iterated Bayesian learning on heterogeneous graphs.

Bayesian language learning on graphs

Represent a population as a set of N learners arranged on a graph. Each learner speaks one of two languages, L_0 or L_1. Population dynamics are included using a birth-death process: at each

time step, a random learner is replaced by a novice learner, the novice learner randomly selects a neighbor, hears an utterance from them, and selects a language based on that utterance. This birth-death process is an abstraction of the biological and cultural processes that shape when and how a learner learns a new language. Although a "birth" may represent an actual birth of a new learner, it might also represent an individual who has chosen to change the language they speak.

Under a Bayesian learning algorithm, learners adopt a language based on a linguistic utterance, D, by selecting a language proportional to the posterior probability of each language,

$$p(L_i|D) = \frac{p(D|L_i)p(L_i)}{p(D|L_0)p(L_0) + p(D|L_1)p(L_1)}. \quad (2)$$

We assume that each utterance is consistent with either L_0 or L_1, and when asked to speak, a teacher correctly produces an utterance consistent with their language with probability $1-\varepsilon$, where ε represents an error rate in production. If an utterance, D, is consistent with a language, L_i, then $p(d|L_i) = 1-\varepsilon$. Innate linguistic preferences are included through the prior probability of each language, $p(L_i)$.

Stationary distribution of languages

In this section, we demonstrate that when learning from a single teacher on heterogeneous graphs, the probability that a specific learner speaks a language after many generations is the same as the prior probability of that language. This extends the result that Griffiths and Kalish (2007) proved for transmission chains to more complex population structures.

An intuition for this result can be obtained by re-imagining the transmission of languages across a graph as a set of chains. In each update, we consider updating the value of a single learner by having that learner learn from a teacher. If we look back in time, that teacher learned their language from someone else, so consider the teacher's teacher. We can then construct a chain of teacher-learner pairs from any individual back to one of the individuals in the initial population. This chain is akin to a transmission chain. The probability that the learner at the end of a chain speaks a language should thus converge to the prior distribution as the chain gets longer.

To make this intuition more precise, we introduce the notion of a Markov process: a process where the probability of future states depends only on the current state. The birth-death process we describe above is a Markov process: each update only depends on the current languages that the learners have adopted, not on the languages spoken by deceased learners. This process is also ergodic: because of the noise in transmission, each learner has a small chance of adopting a different language than their teacher, preventing a certain assignment of languages to learners in the population becoming fixed.

The Markov property allows us to examine the long-term dynamics of language change in this population. Given a population of N learners, let the binary vector s represent that state of learners in the population (the language that each learner speaks). Because this process is Markov, the probability of a future state s_t just depends on the current state, s_{t-1}. This process allows us to define a probability distribution on future outcomes, p_t, where $p_t(s)$ is the probability of s after t time steps. Because this process is ergodic, there exists a stationary distribution, p, over future states defined by $p_t(s) \to p(s)$ as $t \to \infty$. To find the probability that a specific learner, i, adopts a language, L_1 (or alternatively L_0) we marginalize over the language spoken by i in state s by the likelihood of s in the stationary distribution,

$$v_i = \sum_s \delta_{L_1}(s_i) p(s). \tag{3}$$

$\delta_{L_1}(s_i)$ is an indicator function that is 1 if s_i speaks language L_1 and 0 otherwise.

To find this value, we note that the stationary distribution is characterized by its invariance to future time steps; if $p_t(s) = p(s)$ then $p_{t+1}(s) = p(s)$. Since v depends only on p, then v is also invariant to future time steps. Given the transition dynamics described above, we find that $v_i = p(L_1)$ for all i satisfies this requirement, and is unique in this regard. The probability that a given learner speaks L_1 at the stationary distribution is the same as the prior distribution. A complete proof is provided in the Supplementary Materials.

Simulations on heterogeneous graphs

In order to verify the analytic predictions above, we used agent-based simulations to find the stationary distribution of a population on a series of graphs. We found that, on average, the population converged to the prior distribution on each graph.

In each simulation, learners in the population had the option of learning two languages. Each population consisted of 100 learners on an undirected graph. We considered learners living on a complete, small world[1] and scale free graphs[2], as well as two-dimensional lattices. These graphs were chosen as types of graphs that are thought to mimic some of the properties of real world populations (Barabasi & Albert, 1999; Watts & Strogatz, 1998).

At the beginning of each simulation, learners randomly adopted one of the two languages with equal probability. At each time step, a learner was randomly selected from the population and replaced by a new learner. The new learner randomly sampled a linguistic utterance from one of its neighbors and adopted a language using the Bayesian learning algorithm described above. The production error rate was $\varepsilon = .05$. Each generation consisted of 100 time steps, enough so that on average each individual is replaced once.

To examine how the prior distribution changed the long term behavior of the population, we varied the prior on L_1 in .1 increments between .5 and .9. We found that in most simulations the population reached its stationary distribution in 50 generations. We averaged the proportion of learners who spoke each language after 50 generations across 1000 simulations. The results are given in Fig. 1(a). We found that the stationary distribution for each social structure was the same as the prior distribution.

These simulations verify our analytic predictions. However, we also found that for non-complete graphs neighbors were more likely to share a language than predicted by chance. To visualize this phenomenon we ran a series of simulations on a two-dimensional lattice. Fig. 1(b-d) shows a sample result, showing that the population contained a number of large clusters of

[1] Created through reattachment of a neighbor graph (for more details see Watts & Strogatz, 1998). The reattachment probability was .1.

[2] Created through preferential attachment (see Barabasi & Albert, 1999).

language speakers where most of the learners spoke the same language. This suggests that even though the population may not converge on a single language, the distribution of languages in the population is not random; individuals are able to speak to their neighbors.

Capturing correlations among learners

One of the criticisms leveled at iterated learning models is that instead of ending up in a heterogeneous mix of languages at the stationary distribution, real-world populations tend to converge on a single language. Fig. 1(b-d) shows that iterated learning on a lattice converged to a mixture of languages characterized by local clusters where neighbors generally spoke the same language. This finding suggests that introducing population structure might let locally homogeneous populations of learners arise, while still allowing for an overall heterogeneous distribution of languages in the population. This would reduce concerns that at the stationary distribution learners may not be able to speak with their neighbors, and thereby potentially increasing the value that language gives the learner (Smith & Kirby, 2008). To investigate this behavior further we borrow tools from statistical physics developed to analyze a general class of dynamic models, which our Bayesian model is an specific example of, voter models.

Voter models

Voter models are a general framework for analyzing how beliefs diffuse across socially structured populations (Castellano, 2012), and are akin to Moran models, another model that has been used to capture the dynamics of language learners in spatially structured populations (Kalish, 2007). In the standard voter model, the nodes of a graph represent learners. Each learner adopts one of two states. At each time step, a single learner is randomly selected and replaced by a novice learner. The novice learner adopts a state based on the states of its neighbors. Two common learning strategies are selecting the state of the majority, or copying the state of a random neighbor. This process is directly analogous to the model we presented in the previous section, where the learners use a Bayesian learning rule to adopt a new state. Previous analyses of

voter models have demonstrated that population structure can have a substantial effect on both the convergence probabilities and convergence rates (Sood et al., 2008; Castellano, 2012).

While most work on voter models has concentrated on deterministic learning rules (e.g. copy a neighbor without error), Schweitzer and Behera (2009) analyzed a probabilistic model. They showed that in this model, two beliefs could co-exist in a population. Given two states, 0 and 1, the expected rate of change of the proportion of learners with state 1 at time t is given by the differential equation

$$\frac{d}{dt}x_1(t) = \sum_{\sigma}[w(1|0,\sigma)x_{0,\sigma}(t) - w(0|1,\sigma)x_{1,\sigma}(t)], \tag{4}$$

where σ denotes the neighborhood of a point, $w(i|1-i,\sigma)$ the probability of an node in state $1-i$ to adopt state i if the neighborhood of the node is σ, and $x_{i,\sigma}$ the frequency of nodes in state i with neighborhood σ.

The iterated learning model analyzed in the previous section is a special case of this probabilistic voter model. In this case the states of learners represent the languages that those learners adopt, and the update rule $w(i|1-i,\sigma)$ can be computed using Equation 2. Our assumptions about the language learning process also lead us to two equivalences: since the probability of adopting a language does not depend on what state the node was in before, and since each learner must adopt a language, $w(i|i-1,\sigma) = w(i|\sigma)$, and $w(i|\sigma) + w(1-i|\sigma) = 1$.

Learning on heterogeneous graphs

In this section we analyze Equation 4 when learners learn from a single teacher. For convenience, let $1-a$ denote the probability that a learner adopts language L_0 after learning from a teacher who speaks L_0. Let $1-b$ denote the probability that a learner adopts language L_1 after learning from a teacher who speaks L_1. a and b act as error rates in language transmission. Values for a and b corresponding to Bayesian learning are provided in the Supplementary Materials.

Applying this to Equation 4 gives that the rate of change of L_1 learners is

$$\frac{d}{dt}x = a\sum_{m=1}^{M}\sum_{k=0}^{m} x_{\sigma_k^m}(t) + (1-a-b)\sum_{m=1}^{M}\sum_{k=0}^{m}\frac{k}{m}x_{\sigma_k^m}(t) - x, \tag{5}$$

where σ_k^m denotes all nodes with m neighbors (up to a maximum degree of M), k of which have adopted language L_1. After simplifying, the summation is

$$\frac{d}{dt}x = a + (1-a-b)E[f] - x, \tag{6}$$

where $E[f]$ is the frequency that nodes in a neighborhood have state 1. $E[f]$ must be calculated on a graph by graph basis. For degree-regular graphs, in which every node has the same number of edges, $E[f] = x$. This means that for degree-regular graphs the stationary distribution of x is the same as what we found in the previous section,

$$x = \frac{a}{a+b}. \tag{7}$$

We demonstrate through simulations that this is also the stationary distribution for non-degree regular graphs like small world or scale free graphs.

In the Supplementary Materials, we demonstrate that in our formulation of Bayesian learners $\frac{a}{a+b} = p(L_1)$. More generally however, for a given transmission process, the stationary distribution of languages will simply depend on the relative error rates, a and b. Other models of language transmission, potentially including other Bayesian models of language learning, may produce different error rates for a and b and alter the stationary distribution of languages in the population. This replicates the result obtained using a Markov Process, demonstrating that the prior distribution of the languages will be the stationary distribution of languages in the population. Using the tools developed here, we can push further on this result and examine what the average number of pairs of same-language speaking nodes are.

Predicted correlations between pairs of learners

We define $x_{1,1}$ to be the frequency of edges where both learners speak language L_1, and develop a differential equation to express how the frequency of pairs changes over time. If the graph is degree-regular, with each node having degree m, this equation is

$$\frac{d}{dt}x_{1,1} = (1-x)\sum_{k=0}^{m} kw(1|0,\sigma_k^m)x_{\sigma_k^m,0} - x\sum_{k=0}^{m} kw(0|1,\sigma_k^m)x_{\sigma_k^m,1}. \tag{8}$$

By adding in the assumption that learners randomly copy a single teacher, and accurately copy state 0 with probability $1-a$ and state 1 with probability $1-b$, the differential equation can be reduced to

$$\frac{d}{dt}x_{1,1} = m(ax - x_{1,1}) + \frac{1-a-b}{k}E[f^2], \tag{9}$$

where $E[f^2]$ is the squared expectation of the frequency of neighbors that have state 1. As with $E[f]$, this quantity depends on the actual structure of the graph.

We can estimate $x_{1,1}$ by using a *pair approximation*. This approximation places a lower bound on the probability that two nodes share the same state, by assuming that the states of neighbors are uncorrelated. In this approximation we assume that if a central node speaks L_1, then the probability that a given neighbor speaks L_1 can be expressed by $\frac{x_{1,1}}{x}$, and the probability that the neighbor speaks L_0 can be expressed by $\frac{x_{1,0}}{x}$. We track the pair probabilities using $x_{1,1}$, $x_{1,0}$, $x_{0,1}$, and $x_{0,0}$[3]. Using this estimate we can solve Equation 8 to get the equilibrium value of $x_{1,1}$ on degree regular graphs. The details of the solution are provided in the Supplementary Materials. Let $d = 1-a-b$. If d is close to 1, we find that we can aproximate the equilibrium value of the correlation between nodes by

$$x_{1,1} \approx x^2 + \frac{m}{m-1}\frac{x(1-x)}{2d} - \frac{1}{2}x(1-x). \tag{10}$$

From this equation, we have that the average degree of a node affects the correlation between

[3]For the full technical details of pair approximation see (e.g. Schweitzer & Behera, 2009).

nodes; on graphs where nodes have an average low degree, nodes will be more likely to share the same state. This effect disappears as the number of neighbors grows. For certain graphical structures, particularly those with a high clustering coefficient, a measure of how likely two neighbors of a central node are to themselves be neighbors, the correlation between nodes may be higher.

To test the predictions made by the voter model, we ran a series of simulations on small-world, scale-free and complete graphs. Across all simulations the prior distribution was set to $p(L_1) = .6$. Otherwise the simulations were identical to those presented earlier. In Fig. 2(a) we show the rate at which learners converge to the prior distribution. In Fig. 2(b), we show the equilibrium value of $x_{1,1} + x_{0,0}$ for small-world, scale-free, and complete graphs. We found that neighbors in small-world networks and two-dimensional lattices, two networks with high clustering coefficients, a feature of real world networks (Newman & Park, 2003), were more likely to share languages than predicted by the model. This suggests that even though the graphical structure does not influence the stationary distribution of languages in a population as a whole, it may influence the local distribution of languages, leading to clusters of homogeneous language speakers. Depending on the relative error rates in learning and the prior distribution of languages these clusters may not be stable, and may change over time as learners in them adopt new languages. At any time point however, we should expect that learners are more likely to be able to speak with their neighbors than by chance alone.

Conclusion

In this paper we examined how population structure can interact with a learner's inductive biases to influence which languages are produced by iterated learning. We proved that, under our model, the structure of the population plays little role in determining whether a given learner speaks a certain language. By introducing the voter model we were also able to examine how the number of neighbors who shared the same language changed over time, a factor that is important in assessing the value of language (Smith & Kirby, 2008). We found that the structure of the

population greatly impacted how likely pairs of learners were able to communicate with each other. These results extend the results of Griffiths and Kalish (2007) to heterogeneous graphs. More generally, they support the generalizability of theoretical and empirical results produced by iterated learning beyond transmission chains. Based on our findings, a reasonable conjecture is that these results should hold in most, if not all, cases where learners learn from a single teacher.

Further work needs to be done to explore how population structure may impact learners who learn from multiple teachers. Smith (2009) showed that in freely mixing populations, the distribution of learners in the population would not converge to the prior distribution. That result was replicated for a small number of population structures by Stadler (2009). In contrast, Burkett and Griffiths (2010) found that learners who learned multiple languages from multiple teachers also converged to the prior distribution on languages. Past work has shown that for even fairly simple learning rules, the dynamics of learners learning from multiple teachers in structured populations may be far more complex (e.g. Castellano, Muñoz, & Pastor-Satorras, 2009).

The results in this paper shine some light on how simple iterated learning models can be extended to real populations. In particular they provide a way to reconcile the predictions of iterated learning models with the geographic distribution of real languages (e.g. Smith & Kirby, 2008). In our simulations we found that there exist local clusters of speakers who share a language. This provides a way to interpret the stationary distribution of iterated learning models: we expect some proportion of speakers to learn each language, but we don't expect those speakers to be scattered randomly throughout the population. Rather, speakers are preferentially assorted with other speakers of the same language, potentially creating local clusters of homogeneous language learners. We hope that further analyses of this kind can be used to bridge the gap between models we can analyze and models that actually capture the dynamics of language evolution in real populations.

References

Barabasi, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509-512.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial life*, *8*(1), 25–54.

Burkett, D., & Griffiths, T. L. (2010). Iterated learning of multiple languages from multiple teachers. In *The evolution of language: Proceedings of the 8th international conference (evolang8)* (pp. 58–65).

Castellano, C. (2012). Social influence and the dynamics of opinions: the approach of statistical physics. *Managerial and Decision Economics*, *33*(5-6), 311–321.

Castellano, C., Muñoz, M. A., & Pastor-Satorras, R. (2009). Nonlinear q-voter model. *Physical Review E*, *80*(4), 041129.

DeGraff, M. (2001). *Language creation and language change: Creolization, diachrony, and development*. MIT Press.

Griffiths, T., & Kalish, M. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441-480.

Kalish, M. (2007). Iterated learning with selection: convergence to saturation. In *The evolution of language: Proceedings of the 6th international conference (evolang7).*

Kirby, S. (2001). Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *Evolutionary Computation, IEEE Transactions on*, *5*(2), 102–110.

Kirby, S., Dowman, M., & Griffiths, T. L. (2007). Innateness and culture in the evolution of language. *Proceedings of the National Academy of Sciences*, *104*(12), 5241–5245.

Newman, M. E., & Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, *68*(3), 036122.

Nowak, M. (2006). *Evolutionary dynamics: Exploring the equations of life*. Harvard University Press.

Schweitzer, F., & Behera, L. (2009). Nonlinear voter models: the transition from invasion to coexistence. *European Physical Journal B*, *67*, 301-318.

Smith, K. (2009). Iterated learning in populations of Bayesian agents. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Smith, K., & Kirby, S. (2008). Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*(1509), 3591–3603.

Smith, K., Kirby, S., & Brighton, H. (2003). Iterated learning: A framework for the emergence of language. *Artificial life*, *9*(4), 371–386.

Sood, V., Antal, T., & Redner, S. (2008). Voter models on heterogeneous networks. *Physical Review E*, *77*(4), 041121.

Stadler, K. (2009). *Cultural transmission and inductive biases in populations of Bayesian learners.*

Watts, D., & Strogatz, S. (1998, 06 04). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442.

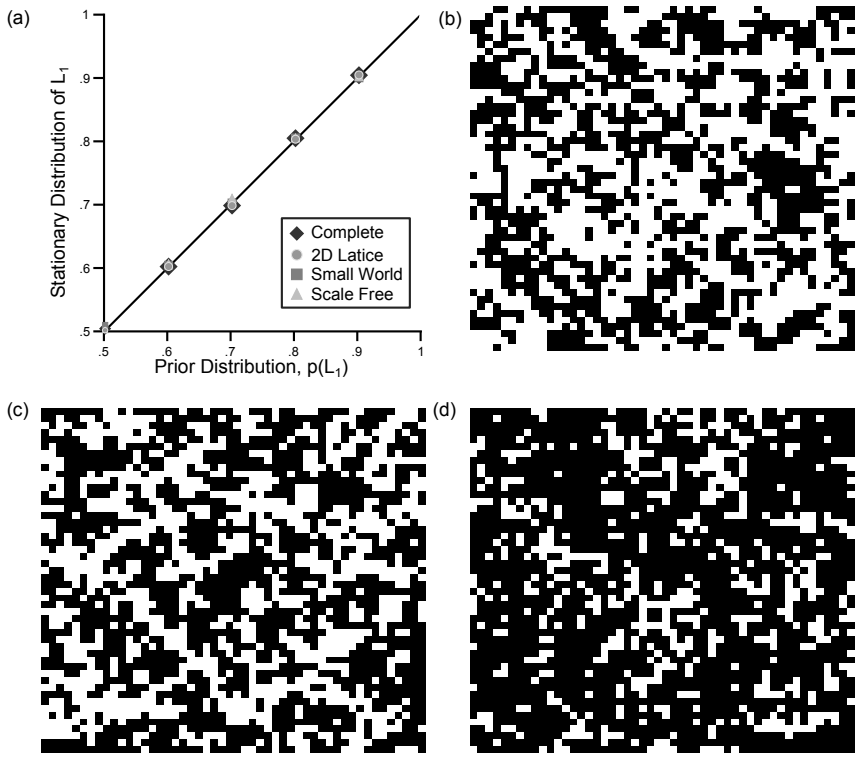

Figure 1. Dynamics of learners on heterogeneous graphs. (a) The stationary distribution for the population as a function of learners' prior beliefs. (b-d) A sample distribution of learners simulated on a 50 by 50 lattice where the prior for L_1 was set to (a) 0.5, (b) 0.6, and (c) 0.8.

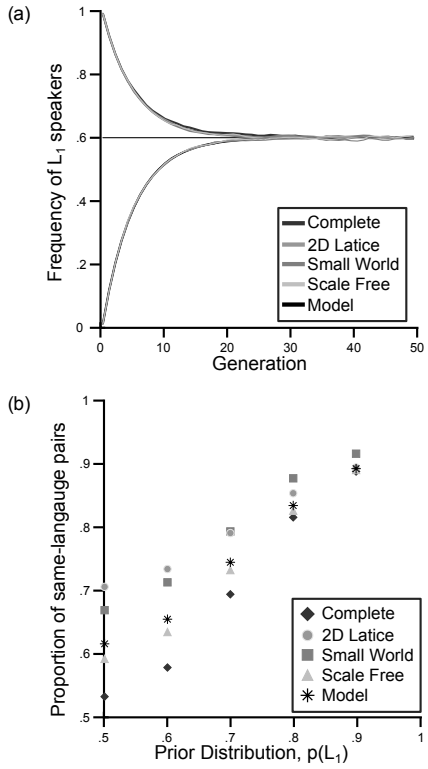

Figure 2. Predictions from the voter model compared with simulations on different types of graphs. (a) Frequency of learners in the population speaking L_1 as a function of the number of generations. (b) Proportion of learner pairs that speak the same language.

Supplementary Materials: Mathematical details

Error rates in Bayesian learning

In this section we derive the probability that a novice learner will adopt language L_0 given that their teacher holds language L_0. Following the setup of Equation **??** we have that if the learner hears an utterance, u, consistent with L_0 the probability that they adopt L_0 is

$$p(L_0|u) = \frac{(1-\epsilon)p(L_0)}{(1-\epsilon)p(L_0) + \epsilon p(L_1)}. \tag{S.1}$$

On the other hand if learners hear an utterance consistent with L_1 the probability that they adopt L_0 is

$$p(L_1|u) = \frac{\epsilon p(L_0)}{\epsilon p(L_0) + (1-\epsilon)p(L_1)}. \tag{S.2}$$

If their teacher speaks language L_0 then the probability of creating an utterance consistent with L_0 is $(1-\epsilon)$ and the probability of creating an utterance inconsistent with L_0 is ϵ. This gives that the probability of adopting L_0 from a teacher who speaks L_0 is

$$(1-\epsilon)\frac{(1-\epsilon)p(L_0)}{(1-\epsilon)p(L_0) + \epsilon p(L_1)} + \epsilon\frac{\epsilon p(L_0)}{\epsilon p(L_0) + (1-\epsilon)p(L_1)}. \tag{S.3}$$

A similar equation can be developed for the probability of learning L_1 from an L_1 language speaker. These equations produce the error rates a and b used in the main text.

Proof of the stationary distribution

In this section we demonstrate that the probability that a given learner in our model after many generations speaks a language is the same as the prior probability of that language. Following the setup in the text, let $s = \{s_1, s_2, ..., s_n\}$ be a binary vector representing a given assignment of beliefs to nodes. Assume that the graph is path connected (Diestel, 2012). Consider a probability distribution over states at time t, p_t. We can examine the probability that the individual node i speaks language L_1 by

marginalizing over possible states,

$$v_{t,i} = \sum_s \delta_{L_1}(s_i) p_t(s). \tag{S.4}$$

where $\delta_{L_1}(s_i)$ is 1 if s_i speaks L_1 and 0 otherwise. The vector v_t expresses the probability that a given node speaks L_1 at time t. Since v_t is a linear combination of elements of p_t, at the stationary distribution of the Markov process, v_t will converge to a value that is invariant to future time steps. We label the value of v_t at the stationary distribution, v. We demonstrate below that $v_i = \frac{a}{a+b}$ for all i by showing that it is invariant to changes in the update, and is unique in this regard.

We first demonstrate its stability. Let $1-a$ be the error rate in learning L_0 from an L_0 speaker, and let $1-b$ be the rate of learning L_1 from a L_1 speaker. Let $v_{t,i} = \frac{a}{a+b}$. We have that

$$v_{t+1,i} = \sum_s \delta_{L_1}(s_i) p_{t+1}(s). \tag{S.5}$$

The state of the node s_i is given by marginalizing over all of the possible actions that could have happened in the transition between $t \to t+1$. With probability $\frac{n-1}{n}$ the node was not changed, and so the value was not changed, $v_{t+1,i} = v_{t,i}$. With probability $\frac{1}{n}$ the value of the node changes. If it changes, then its new value depends on the value of its m neighbors indexed by $\{a_1, ...a_m\}$ at time t:

$$\begin{aligned} v_{t+1,i} = \sum_s \frac{1}{m} \sum_{j=0}^{m} [&(1-b)\delta_{L_1}(s_{t,a_j}) p_t(s) \\ &+ a(1-\delta_{L_1}(s_{t,a_j})) p_t(s)], \end{aligned} \tag{S.6}$$

We also have $\sum_s \delta_{L_1}(s_{t,a_j})) p_t(s) = v_{a_j,t}$ and $\sum_s p_t(s) = 1$. Under the assumption that

$v_i = \frac{a}{a+b}$, this produces

$$v_{t+1,i} = \frac{1}{m}\sum_{j=0}^{m}(1-b)v_{a_j,t} + a(1-v_{a_j,t}) \tag{S.7}$$

$$= (1-b)\frac{a}{a+b} + a\frac{b}{a+b} \tag{S.8}$$

$$= \frac{a}{a+b} = v_{t,i}. \tag{S.9}$$

Even if the node is updated, v_t does not, on average change; our choice for v is invariant to future time changes.

Moreover v is unique in this regard. Suppose there existed another distribution w_t. If $w_i = w_j$ for all j, i then Equation S.7 simplifies to

$$w_i = (1-b)w_i + a(w_i) \tag{S.10}$$

which only has a single fixed point, $w_i = v_i$. Suppose instead that there exists $w_i \neq w_j$, then Equation S.7 gives us that after an update

$$w_i = (1-b)\sum \frac{w_{a_j}}{m} + a\sum \frac{1-w_{a_j}}{m} \tag{S.11}$$

where the sum is over the m neighbors of w_i, indexed by a_j. Choose the largest value of w_i such that for one of its neighbors, $w_{a_j} < w_i$ (such a neighbor exists since the graph is path connected). Suppose $w_i > \frac{a}{a+b}$. We have that $w_i > \sum \frac{w_{a_i}}{m}$. Let $f(x) = (1-b)x + a(1-x)$, either $f(x) < \frac{a}{a+b}$ or $f(x) < x$. In either case, after the update, $f(\sum \frac{w_{a_i}}{m}) < w_i$, implying that w is not invariant to updates. If $w_i < \frac{a}{a+b}$ consider instead the smallest value of w_i such that for one of its neighbors, $w_{a_j} > w_i$, and demonstrate that it must increase after the update.

This demonstrates that v is both invariant under updates and is unique; v_i represents the probability that each node is in state i at the stationary distribution. This gives that the likelihood that any individual learner speaks L_1 is $\frac{a}{a+b}$.

We now demonstrate that $\frac{a}{a+b} = p(L_1)$ for Bayesian learners. From Equation S.3 a is given by

$$\epsilon \frac{(1-\epsilon)p(L_1)}{(1-\epsilon)p(L_0)+\epsilon p(L_1)} + (1-\epsilon)\frac{\epsilon p(L_1)}{\epsilon p(L_0)+(1-\epsilon)p(L_1)}.$$

and b is given by

$$(1-\epsilon)\frac{\epsilon p(L_0)}{(1-\epsilon)p(L_0)+\epsilon p(L_1)} + \epsilon \frac{(1-\epsilon)p(L_0)}{\epsilon p(L_0)+(1-\epsilon)p(L_1)}.$$

We have that, $b = \frac{p(L_1)}{p(L_0)}a$. Since there are only two languages in the population we can set $p(L_0) + p(L_1) = 1$. This gives that $b = \frac{1-p(L_1)}{p(L_1)}a$. Applying this to Equation **??** gives

$$\frac{a}{a+b} = \frac{a}{a+\frac{1-p(L_1)}{p(L_1)}a} = \frac{1}{1+\frac{1}{p(L_1)}-1} = p(L_1). \tag{S.12}$$

Learning on heterogeneous graphs

In this section we derive Equation **??** from Equation **??**. Equation **??** gives that

$$\frac{d}{dt}x_1(t) = \sum_{\sigma}[w(1|0,\sigma)x_{0,\sigma}(t) - w(0|1,\sigma)x_{1,\sigma}(t)].$$

Denote σ_k^m to be the set of all neighborhoods with $|\sigma_k^m| = m$ and k nodes with state 1. Let M be the maximum degree of any node on the graph. Since the population is finite, the sum is well defined. We can rewrite Equation **??** as

$$\frac{d}{dt}x_1(t) = \sum_{m=1}^{M}\sum_{k=0}^{m}[w(1|\sigma_k^m)x_{0,\sigma_k^m}(t) - w(0|\sigma_k^m)x_{1,\sigma_k^m}(t)].$$

Since $w(0|\sigma) = 1 - w(1|\sigma)$, and $x_{1,\sigma_k^m}(t) + x_{0,\sigma_k^m}(t) = x_{\sigma_k^m}(t)$ then

$$\frac{d}{dt}x_1(t) = \sum_{m=1}^{M}\sum_{k=0}^{m}[w(1|\sigma_k^m)x_{\sigma_k^m}(t) - x_{1,\sigma_k^m}(t)]$$

Note that $\sum_{m=1}^{M}\sum_{k=0}^{m} x_{1,\sigma_k^m}(t) = x_1(t)$, so we can break up the summation to obtain

$$\frac{d}{dt}x_1(t) = \sum_{m=1}^{M}\sum_{k=0}^{m} w(1|\sigma_k^m)x_{\sigma_k^m}(t) - x_1(t). \tag{S.13}$$

For learning from a single teacher, where the probability in copying state 0 is $1-a$ and the probability of copying state 1 is $1-b$, let m denote the degree of a node in question and $\frac{k}{m}$ the proportion of neighbors in state 1. Then

$$w(1|\sigma_k^m) = (1-b)\frac{k}{m} + a\frac{m-k}{m} = a + \frac{k}{m}(1-a-b)$$

Let x denote $x_1(t)$. Applying this to Equation S.13 gives,

$$\frac{d}{dt}x = \sum_{m=1}^{M}\sum_{k=0}^{m}[a + \frac{k}{m}(1-a-b)]x_{\sigma_k^m}(t) - x.$$

Breaking up the summation gives that

$$\frac{d}{dt}x = a\sum_{m=1}^{M}\sum_{k=0}^{m}x_{\sigma_k^m}(t) + (1-a-b)\sum_{m=1}^{M}\sum_{k=0}^{m}\frac{k}{m}x_{\sigma_k^m}(t) - x.$$

We have that $\sum_{m=1}^{M}\sum_{k=0}^{m}x_{\sigma_k^m}(t) = 1$ and $\sum_{m=1}^{M}\sum_{k=0}^{m}\frac{k}{m}x_{\sigma_k^m}(t) = E[f]$, where f is the frequency of nodes in a neighborhood who have state 1. The equilibrium mean is then given by the solution to

$$0 = a + (1-a-b)E[f] - x.$$

For degree regular graphs, let η_i denote the state of node i and N_i the neighborhood of i then

$$E[f] = \frac{1}{N}\sum_{i=1}^{N}\sum_{j\in N_i}\frac{\eta_i}{deg(j)}.$$

If all the nodes have the same degree then $|N_i| = \deg(j)$ for all $j \in N_i$. This gives that $E[f] = x$, and the equilibrium expectation is

$$x = \frac{a}{a+b}. \tag{S.14}$$

Approximating the correlations between nodes

In this section we estimate the correlations between the states of neighboring nodes when the population is at the stationary distribution. From Equation **??** we are interested in the fixed point of

$$\frac{d}{dt}x_{1,1} = m(ax - x_{1,1}) + \frac{d}{m}E[f^2].$$

We can break up the squared expectation term by looking at the squared expectation for the neighborhoods of nodes in state 0, $E[f_0^2]$, and in state 1, $E[f_1^2]$.

For state 1, the number of nodes in the neighborhood that also have a state of 1 is approximated by the pair approximation of having been drawn from a binomial distribution, with parameter $\frac{x_{1,1}}{x}$. This gives that the squared expectation can be found as a difference between the square of the expectation and the variance of a binomial distribution. We have that

$$E[f_1^2] = Var(f_1^2) + E[f_1]^2 = \frac{m(m-1)x_{1,1}^2}{x} + mx_{1,1}.$$

A similar equation can be derived for f_0. Substituting this back into the previous expression returns an equation, quadratic in $x_{1,1}$. Let $d = 1 - a - b$ then the equation is

$$\begin{aligned}\frac{d}{dt}x_{1,1} = x_{1,1}^2\frac{d(m-1)}{x(1-x)} + x_{1,1}(-m - \frac{2xd(m-1)}{1-x}) \\ + \frac{x^2d(m-1)}{1-x} + dx + a(xm)\end{aligned}$$

The zeroes of this equation can be found via application of the quadratic equation:

$$x_{1,1} = x^2 + \frac{x - x^2}{2(d - d/m)} + g(x)$$

where $g(x)$ is

$$\sqrt{(x^2 + \frac{x-x^2}{2(d-d/m)})^2 - \frac{a(x^2-x^3)}{d-d/m} - \frac{x^2-x^3}{m-1} + x^3}.$$

We found that for our simulations $g(x)$ can be reasonably approximated by $\frac{1}{2}x(1-x)$.

References

Diestel, R. (2012). *Graph theory, 4th edition* (Vol. 173). Springer.